PerFairX: Is There a Balance Between Fairness and Personality in Large Language Model Recommendations?

Chandan Kumar Sah
School of Computer Science and Engineering
Beihang University
Haidian, Beijing, China
sahchandan98@buaa.edu.cn

Abstract

The integration of Large Language Models (LLMs) into recommender systems has enabled zero-shot, personality-based personalization through prompt-based interactions, offering a new paradigm for user-centric recommendations. However, incorporating user personality traits via the OCEAN model highlights a critical tension between achieving psychological alignment and ensuring demographic fairness. To address this, we propose PerFairX, a unified evaluation framework designed to quantify the trade-offs between personalization and demographic equity in LLM-generated recommendations. Using neutral and personality-sensitive prompts across diverse user profiles, we benchmark two state-of-the-art LLMs, ChatGPT and DeepSeek, on movie (MovieLens 10M) and music (Last.fm 360K) datasets. Our results reveal that personality-aware prompting significantly improves alignment with individual traits but can exacerbate fairness disparities across demographic groups. Specifically, DeepSeek achieves stronger psychological fit but exhibits higher sensitivity to prompt variations, while ChatGPT delivers stable yet less personalized outputs. PerFairX provides a principled benchmark to guide the development of LLM-based recommender systems that are both equitable and psychologically informed, contributing to the creation of inclusive, user-centric AI applications in continual learning contexts.

1. Introduction

In recent years, recommender systems have entered a new era with the incorporation of large language models (LLMs) as core components. LLM-based recommenders powered by models like GPT-4 can interpret rich contextual prompts and generate item suggestions in natural language, enabling more conversational and flexible recommendations. Early studies suggest that LLMs even possess promising zero-shot capabilities, producing meaningful recommendations without task-specific training [26]. However, the adoption of LLMs in recommendation brings new challenges in **fairness** and prompt reliability. Fairness in recommender systems is inherently multifaceted: unlike simple classification, recommendations involve multiple stakeholders, ranked outputs, and heavy personalization, complicating how we define and measure fairness [5, 6, 40]. These complexities are now compounded by LLMs' known biases and sensitivity to prompts. Recent observations show that LLM-generated recommendations can reflect unfair biases; for example, ChatGPT's movie suggestions for users from the African continent were systematically worse than for other users [45]. Such disparities mirror real-world social biases that a recommender should not perpetuate. In addition, RecLLMs exhibit high *prompt sensitivity* minor rephrasings of the input can yield significantly different outputs. An LLM may suggest entirely different items if a user profile is worded slightly differently, as the model can be unduly swayed by prompt wording or the order in which informa-

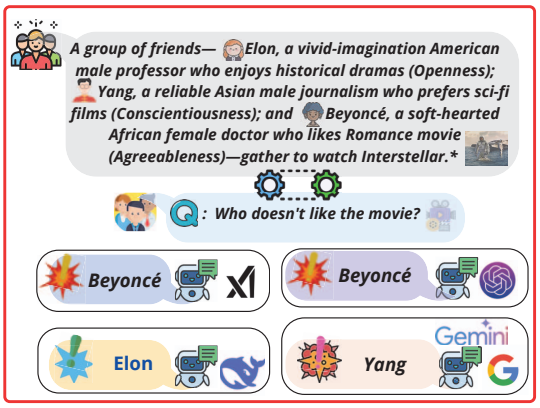

Figure 1. An example of personality-driven movie preferences in a diverse group, examining fairness in LLM recommendations.

[0]This paper includes examples that may be offensive, harmful, or biased, used solely for academic research purposes.

tion is presented [8]. Although careful prompt engineering can reduce some of these effects [10], ensuring stable and fair outputs remains a pressing concern.

Missing from current literature is a focus on what we term **psychological fairness**. Most fairness research in recommender systems (including recent RecLLM studies) has centered on demographic attributes, often using interventions (e.g., toggling a user's stated gender or location) to detect bias [11–13, 33]. However, this may conflate genuine preference differences with unfair discrimination. One such underexplored source of variation is user personality. Meanwhile, a separate line of work on personality-aware recommenders has shown that incorporating users' psychological traits can improve recommendation relevance and user satisfaction. [11, 19, 36, 47]. The Big Five personality model (OCEAN) is commonly used to represent key trait dimensions [34]. Yet, no prior study has jointly examined fairness and personality alignment in recommender systems. It remains unknown whether adapting recommendations to user personalities might inadvertently introduce new biases, or if such personalization can be achieved without sacrificing fairness. To illustrate this tension, we conducted a case study (Figure 1) with three personas and based on Human evaluators ($n = 120$, recruited via LinkedIn, Facebook, X, and WeChat) overwhelmingly identified Elon (80%) as the least likely to enjoy the film. Yet, four leading LLMs overlooked this mismatch, prioritizing genre-balancing heuristics over personality alignment. This divergence highlights systemic biases in LLM recommendations, where surface-level fairness (e.g., equal genre representation) fails to account for psychological congruence. In this work, we aim to fill the gap in understanding how fairness and personality alignment interact in LLM-based recommendation systems. We introduce PerFairX, a benchmarking framework designed to jointly evaluate demographic fairness and psychographic personalization. Our approach leverages prompt engineering to encode user personality traits—based on the OCEAN model—into both neutral and personality-sensitive prompts. We conduct experiments across two domains. For each domain, we simulate personality vectors from user behavior and prompt two leading LLMs—OpenAI's ChatGPT and DeepSeek—to generate top-K recommendations. The results are evaluated using ten metrics encompassing fairness, personality fit, diversity, and accuracy. Our findings reveal that user perceptions of fairness, trust, and satisfaction differ significantly across models and prompts, especially under personality-sensitive settings. To our knowledge, PerFairX is the first framework to systematically explore this dual lens of fairness and personalization in RecLLMs. **Contributions.** Our contributions are as follows: (i) We introduce a psychological fairness perspective into recommender systems by integrating user personality traits, via the OCEAN model, into fairness evaluation, extending traditional demographic-focused audits to address an underexplored dimension of human-centered personalization; (ii) We develop PerFairX, a structured framework that combines prompt engineering, OCEAN-based personality profiling, and ten metrics targeting fairness and alignment, enabling a principled, multi-dimensional assessment of LLM-based recommendations across diverse prompt conditions; and (iii) We perform an empirical evaluation of two leading LLMs ChatGPT and DeepSeek using neutral and personality-sensitive prompts, uncovering distinct model behaviors and highlighting tensions between personality alignment, demographic fairness, and recommendation accuracy.

2. Related Work

The integration of large language models into recommendation systems has generated significant interest, particularly in evaluating fairness in LLM-generated recommendations [10, 23, 45]. While prior work extensively explores group and demographic fairness in traditional RecSys [30, 40], fairness within LLM-based recommendations remains nascent and fragmented. The FaiRLLM benchmark [45] established foundational metrics for evaluating user-side fairness in RecLLMs by examining recommendation discrepancies across sensitive attributes. Building on this, CFaiRLLM [10] introduces intersectional fairness and user-profile sampling, highlighting the complexity of fairness when dealing with multi-attribute user groups. Other efforts like IFairLRS [18] and FairEvalLLM [9] further refine fairness analysis through user-side and item-side lenses, uncovering biases tied to gender, age, and cultural contexts. However, these works overlook psychological factors such as personality traits. Personality modeling, particularly using the OCEAN framework [34], has proven valuable in personalized recommendation and fairness-aware profiling [17, 22, 31, 41]. Few studies, however, combine both fairness and personality dimensions in the context of LLMs. On the prompt engineering front, several studies explore how prompt sensitivity impacts LLM-based RecSys outcomes [7, 15, 16, 21, 48]. While they demonstrate that prompt formulation introduces content and group-level biases, no work systematically investigates how personality-aware prompt variations affect fairness, particularly using LLM APIs like ChatGPT or DeepSeek. Recent work on empathetic recommendations [35, 46] and group fairness in RecLLMs [37] suggests emerging interest in combining user psychology and fairness evaluation. Yet, these studies lack structured benchmarks and composite evaluation metrics integrating demographic and psychological fairness. Our PerFairX framework fills this gap by introducing personality-based fairness alongside traditional fairness and prompt sensitivity measures, contributing a unified methodology for balancing personalization and equity in LLM-powered recommendations.

3. PerFairX Framework

3.1. Personality-Aligned Fairness Evaluation in LLM Recommendations (PAFER)

3.1.1. Fairness Definition.

In the context of large language model-based recommender systems (RecLLMs), we define fairness as the absence of systemic prejudice or favoritism toward user groups defined by sensitive attributes (e.g., gender, age, continent, occupation) in the generation of recommendations. Our focus is on user-side fairness, which ensures that users with comparable preferences and personalities receive equitable treatment—regardless of their demographic group especially when such sensitive information is not explicitly present in the prompt. Furthermore, this work extends fairness analysis beyond traditional demographic parity to incorporate personality-aware fairness. We examine whether recommendations are equitably aligned with users' inferred psychographic traits (e.g., OCEAN dimensions), ensuring that LLMs do not inadvertently prioritize or marginalize certain personality types during generation. This dual demographic and psychographic view forms the basis of our proposed PerFairX framework.

3.1.2. Evaluation Overview.

Our framework, **PerFairX**, provides a modular pipeline for evaluating fairness and personality alignment in LLM-based recommendation systems. LLM querying of both ChatGPT[1] and DeepSeek[2] APIs are queried using neutral and personality-sensitive prompts to generate top-K recommendations. The outputs are evaluated using a comprehensive set of metrics spanning personality fit (PAS, GPA), fairness (DP, EO, ILF), prompt sensitivity (SNSR@K, SNSV@K, Jaccard@K), and standard accuracy measures (Precision@K, Recall@K), enabling multi-faceted analysis of recommendation quality and equity. We assess two LLMs across movie and music domains using the MovieLens 10M and Last.fm 360K datasets. For each user, we infer Big Five (OCEAN) personality traits through genre-affinity and behavioral signals, and generate two prompts—one neutral and one personality-sensitive. This setup enables a controlled comparison of how LLMs respond to generic versus personalized queries and allows us to systematically examine fairness, alignment, diversity, and accuracy across demographic and psychological user attributes.

3.1.3. Benchmark Metrics.

We employ 10 key evaluation metrics to assess the fairness, personality alignment, and quality of LLM-based recommendations. These metrics are grouped into three categories:

Personality Fit. Personality fit is assessed using Personality Alignment Score (PAS) and Genre-Personality Alignment (GPA), which together capture how well recommendations reflect the user's inferred psychological profile.

- **Personality Alignment Score (PAS)**: Measures the cosine similarity between the user's OCEAN vector $\vec{p}_u$ and the inferred genre vector $\vec{g}_u$ from the LLM's recommended items.

$$PAS(u) = \frac{\vec{p}_u \cdot \vec{g}_u}{\|\vec{p}_u\| \cdot \|\vec{g}_u\|} \tag{1}$$

  where $\vec{p}_u$ represents the OCEAN personality vector for user u, and $\vec{g}_u$ is the genre vector inferred from the LLM's recommendation. The dot product between these vectors captures how similar the user's personality is to the recommended genres, normalized by the magnitude of each vector[20].
- **Genre-Personality Alignment (GPA)**: Aggregates weighted overlaps between recommended genres and those linked to the user's dominant OCEAN traits.

$$GPA(u) = \sum_{g \in G_{rec}} \sum_{t \in OCEAN} \mathbb{I}_{g \in G_t} \cdot p_u^t \tag{2}$$

  where G_{rec} is the set of genres recommended to user u, and G_t denotes the set of genres associated with each of the OCEAN traits $t \in OCEAN$. The indicator function $\mathbb{I}_{g \in G_t}$ is 1 if genre g aligns with the trait t, and p_u^t is the user's score for that trait[1].

Fairness. We evaluate fairness using Demographic Parity (DP) and Equal Opportunity (EO), which measure group-level disparities, along with Intra-list Fairness (ILF), which captures diversity within individual recommendation lists.

- **DP (Demographic Parity)**: Measures the difference in recommendation probability across sensitive groups.

$$DP = |P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)| \tag{3}$$

  where $\hat{Y}$ is the recommendation outcome (1 if an item is recommended, 0 if not), and A is a sensitive attribute (e.g., gender, race). This metric compares the recommendation probability between two groups (e.g., male and female, or different races[5]).
- **EO (Equal Opportunity)**: Measures the fairness of recommendations given positive outcomes.

$$EO = |P(\hat{Y} = 1|Y = 1, A = 0) - P(\hat{Y} = 1|Y = 1, A = 1)| \tag{4}$$

  where Y is the true relevance (whether the item is actually relevant), and A represents sensitive attributes. This metric compares the true positive recommendation rates between two sensitive groups (e.g., how equally likely a relevant item is recommended for both genders[40]).

[1] https://platform.openai.com/docs/models/gpt-4o
[2] https://api-docs.deepseek.com/

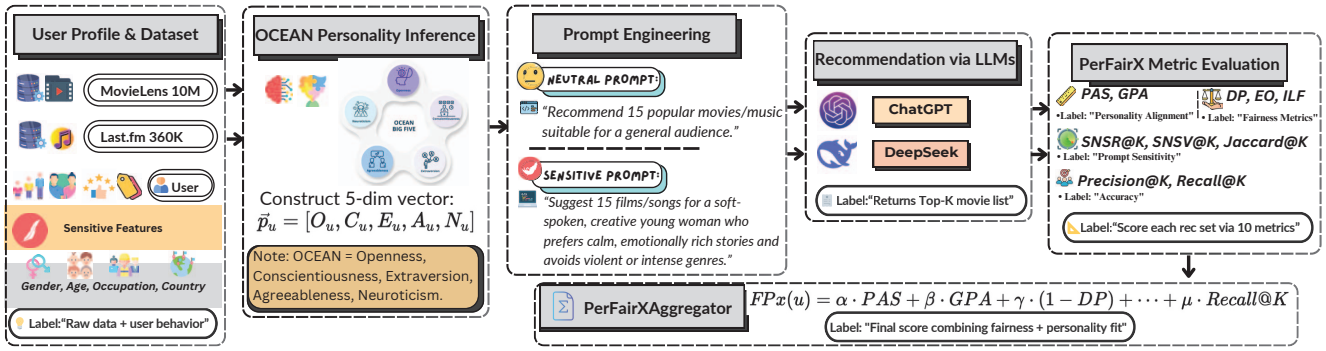

Figure 2. Overview of the PerFairX framework. User data from MovieLens and Last.fm are mapped to OCEAN personality vectors, used to construct neutral and personality-sensitive prompts. LLM-generated recommendations are evaluated across ten metrics, producing a unified PerFairX score.

- **ILF (Intra-list Fairness)**: Measures the diversity of recommendations within a list.

$$ILF@K = -\sum_{g \in G} p(g) \cdot \log p(g) \tag{5}$$

where G is the set of items in the recommendation list, and $p(g)$ is the probability distribution of items. This metric evaluates how evenly items are distributed across different categories, ensuring that the recommendations don't favor certain types of items disproportionately[18].

Prompt Sensitivity. We assess prompt sensitivity using SNSR@K (Sensitive-to-Neutral Similarity Range), SNSV@K (Similarity Variance), and Jaccard@K, which together quantify how much the model's recommendations change in response to personality-sensitive prompting.

- **SNSR@K (Sensitive-to-Neutral Similarity Range) [45]**: Measures the maximum variation in recommendation overlap between sensitive and neutral prompts.

$$SNSR@K = \max_{a \in A} \left| \frac{|R_a^K \cap R_n^K|}{K} \right| - \min_{a \in A} \left| \frac{|R_a^K \cap R_n^K|}{K} \right| \tag{6}$$

where R_a^K and R_n^K represent the top-K recommendations under sensitive and neutral prompts for group a. This metric measures how much the top-K recommendations fluctuate when switching between sensitive and neutral prompts.

- **SNSV@K (Sensitive-to-Neutral Similarity Variance)**: Measures the variance in recommendation overlap across sensitive groups.

$$SNSV@K = \mathrm{Var}_{a \in A} \left(\frac{|R_a^K \cap R_n^K|}{K} \right) \tag{7}$$

This metric calculates the variance in the recommendation overlap for sensitive groups under neutral and sensitive prompts. Higher variance suggests instability in the model's recommendations based on how the user profile is framed[45].

- **Jaccard@K[29]**: Measures the similarity of the top-K recommendations between neutral and sensitive prompts.

$$Jaccard@K = \frac{|R_{neutral}^K \cap R_{sensitive}^K|}{|R_{neutral}^K \cup R_{sensitive}^K|} \tag{8}$$

This metric calculates the Jaccard similarity, comparing the overlap of recommendations under neutral vs. sensitive prompts. A higher Jaccard score suggests that the prompt variation does not heavily impact the recommendations.

Standard Metrics. We use Precision@K and Recall@K to evaluate the relevance and retrieval effectiveness of the top-K recommendations, providing a baseline for assessing overall recommendation quality.

- **Precision@K[26]**: Measures the proportion of top-K recommendations that are relevant.

$$Precision@K = \frac{|Rel \cap Rec@K|}{|Rec@K|} \tag{9}$$

where Rel is the set of relevant items, and $Rec@K$ is the top-K recommended items. This metric measures the accuracy of the top-K recommendations by calculating the proportion that is relevant.

- **Recall@K[8]**: Measures the proportion of relevant items in the top-K recommendations.

$$Recall@K = \frac{|Rel \cap Rec@K|}{|Rel|} \tag{10}$$

This metric captures how many of the relevant items are retrieved in the top-K list. It ensures that the recommender system isn't missing out on relevant items, even if they are not ranked at the top.

3.2. FPx Aggregator

To summarize performance, we define the FPx score:

$$\begin{aligned}\text{FPx}(u) = \alpha \cdot \text{PAS} + \beta \cdot \text{GPA} + \gamma \cdot (1 - \text{DP}) \\ + \delta \cdot (1 - \text{EO}) + \epsilon \cdot \text{ILF} + \zeta \cdot \text{Jaccard@K} \\ + \eta \cdot \text{Precision@K} + \mu \cdot \text{Recall@K}\end{aligned} \tag{11}$$

This scalar score reflects both user alignment and fairness, supporting model comparisons under personality-aware settings.

3.3. Dataset and Preprocessing

We evaluate our PerFairX framework on two large-scale public datasets: **MovieLens 10M** and **Last.fm 360K**, which provide complementary domains (movies vs. music) for assessing fairness and personalization in LLM-based recommendations. Table 1 summarizes key statistics and sensitive features for each dataset. For both datasets, we perform the following preprocessing steps:

- **User Filtering:** We retain users with at least 200 interactions to ensure reliable personality inference and meaningful LLM-based evaluations.
- **Trait Inference via Genre Mapping:** Building on prior work in psychometric personalization, we map movie and music genres to the Big Five (OCEAN) traits. For example, genres such as Sci-Fi and Documentary are linked to high Openness, while Romance may correlate with Agreeableness. This mapping allows us to simulate personality profiles and evaluate alignment.

3.4. Personality Modeling.

We represent each user u as a five-dimensional vector grounded in the Big Five personality traits (OCEAN model), denoted by:

$$\vec{p}_u = [O_u, C_u, E_u, A_u, N_u] \in [0, 1]^5$$

where each dimension corresponds to Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Inspired by McCrae and John [27], we infer these traits using behavioral proxies extracted from user interaction data such as genre affinity distributions, rating dispersion, temporal activity, and catalog diversity.

To strengthen psychological grounding, we integrate trait-genre mappings aligned with prior work in music preference modeling [14]. Specifically, traits like Openness correlate with exploratory genres (e.g., Indie, Jazz), while traits like Conscientiousness align with structured content (e.g., Documentaries, Classical). These mappings help simulate a user's cognitive-emotional profile and enable trait-sensitive prompt generation and fairness evaluation.

3.5. Prompt Design.

We design two types of prompts to evaluate the impact of personality-sensitive conditioning on LLM-generated recommendations: a generic neutral prompt and a personality-aligned sensitive prompt. This prompt distinction is central to our framework and supports comparative evaluation of fairness and alignment.

Neutral Prompt: "Please recommend 15 popular movies/music suitable for a general audience."

Sensitive Prompt Example (for an Introverted User): "I am an introverted movie lover who prefers thoughtful, emotional stories. Please recommend 15 movies/music."

These prompts reflect the contrast between system-wide generalization and individualized personalization. The *neutral prompt* provides a baseline for group fairness, whereas the *sensitive prompt* aims to reflect users' psychological traits (e.g., introversion, agreeableness) inferred from their content preferences. This design choice allows us to evaluate whether personality-aware prompting leads to noticeable behavioral shifts in recommendations and to what extent these shifts remain equitable across demographic and personality subgroups. Figure 2 illustrates how these prompt types are embedded within the PerFairX evaluation pipeline.

4. Results and Analysis

In this section, we present our evaluation of the **PerFairX** framework addressing three research questions. We assess two LLM-based recommendation systems, ChatGPT and DeepSeek, across two domains: movies and music. For the movie domain, we use 5 representative user profiles from the **MovieLens 10M** dataset. For the music domain, we select 5 additional profiles from the **Last.fm 360K** dataset, each representing diverse listening behaviors and inferred personality traits. For every user, recommendations were generated using both neutral and personality-sensitive prompts. The top-15 outputs were evaluated using 10 metrics: Personality Alignment Score (PAS), Genre-Personality Alignment (GPA), Demographic Parity (DP), Equal Opportunity (EO), Intra-list Fairness (ILF), Semantic Novelty Similarity Ratio (SNSR@15), Semantic Novelty Sensitivity Variation (SNSV@15), Jaccard@15, Precision@15, and Recall@15. In addition, we analyze fairness across several sensitive demographic attributes, including gender, age, and occupation. Key quantitative findings are presented in Tables 2, 3, and 4, while qualitative insights and visual analyses are illustrated in Figures 3, 4, and 5.

Research Questions

RQ1: To what extent do LLM-based recommendation systems align with inferred user personality traits (OCEAN model) under different prompt conditions?, **RQ2**: How

Dataset	Female/Male	Senior/Young	Items	Records	Ratings	Sensitive Features
MovieLens 10M	17.4k / 43.6k	1.4k / 4.6k	10.0k	10,000.1k	1–5 stars	⚥ Gender, 🎂 Age, 💼 Occupation
Last.fm 360K	2.0k / 8.0k	1.3k / 8.7k	186.6k	598.0k	Implicit (Plays)	⚥ Gender, 🎂 Age, 🌐 Country

Table 1. Comparison of Key Dataset Statistics for MovieLens 10M and Last.fm 360K.

Table 2. Personality alignment metrics (PAS, GPA) by model, prompt condition, and dataset. Higher values indicate better alignment.

Dataset	Model	Prompt	PAS	GPA
MovieLens	ChatGPT	Neutral	0.749	0.709
	ChatGPT	Sensitive	0.739	0.407
	DeepSeek	Neutral	0.280	0.713
	DeepSeek	Sensitive	0.848	0.336
Last.fm	ChatGPT	Neutral	0.782	0.695
	ChatGPT	Sensitive	0.728	0.421
	DeepSeek	Neutral	0.315	0.731
	DeepSeek	Sensitive	0.872	0.352

Table 3. Fairness metrics (DP, EO) and intra-list fairness (ILF) across datasets. Lower DP/EO indicates better demographic fairness; higher ILF indicates greater diversity.

Dataset	Model	Prompt	DP	EO	ILF
MovieLens	ChatGPT	Neutral	0.679	0.762	0.033
	ChatGPT	Sensitive	0.825	0.952	0.310
	DeepSeek	Neutral	0.237	0.391	0.475
	DeepSeek	Sensitive	0.726	0.901	0.968
Last.fm	ChatGPT	Neutral	0.642	0.728	0.045
	ChatGPT	Sensitive	0.801	0.935	0.289
	DeepSeek	Neutral	0.255	0.377	0.502
	DeepSeek	Sensitive	0.711	0.884	0.932

do prompt variations influence fairness across demographic groups, and what is the trade-off with personality fit?, **RQ3**: How do different LLMs compare in balancing fairness and personality in recommendations?

4.1. RQ1: Personality Alignment

We first investigate how well the generated recommendations align with users' inferred Big Five personality traits. The **Personality Alignment Score (PAS)** quantifies the cosine similarity between a user's OCEAN vector and the latent traits of recommended items. In parallel, the **Genre–Personality Alignment (GPA)** evaluates whether the genres in the recommendations correspond with those typically preferred by users possessing certain dominant personality traits. Table 2 presents the PAS and GPA scores for ChatGPT and DeepSeek under both neutral and personality-sensitive prompts across the MovieLens and Last.fm datasets. Notably, DeepSeek demonstrates a substantial improvement in PAS when personality-sensitive prompting is applied (from 0.280 to 0.848 on MovieLens, and from 0.315 to 0.872 on Last.fm). This highlights its strong adaptability to psychographic cues. By contrast, ChatGPT's PAS remains stable across prompt types. However, GPA scores tend to decrease under sensitive prompts for both models, suggesting that while personality alignment improves, genre alignment weakens. This trade-off reflects the challenge of optimizing for both personality and genre relevance simultaneously. The confirmation of the observation in Table 2 of these results suggests that personality-sensitive prompts significantly improve personality alignment—particularly for DeepSeek—though often at the expense of genre alignment, highlighting the inherent trade-off in personalized recommendation generation.

4.2. RQ2: Fairness vs. Personality Trade-off

In this section, we investigate how prompt variations influence fairness across demographic groups and the trade-off between fairness and personality fit. Specifically, we compare neutral vs. personality-sensitive prompts in terms of three fairness-related metrics: Demographic Parity (DP), Equal Opportunity (EO), and Intra-List Fairness (ILF). While DP and EO assess disparities in exposure and success rates across sensitive groups, ILF measures diversity within the recommendation list.

As summarized in Table 3, the use of personality-sensitive prompts leads to increased demographic disparity (higher DP and EO), but simultaneously improves ILF, suggesting that personalization improves diversity at the cost of fairness. For instance, DeepSeek achieves an ILF of 0.968 on MovieLens under the sensitive condition, up from 0.475 in the neutral prompt, but with a corresponding increase in DP and EO—indicating a rise in group-level bias. Figure 4 visualizes the fairness-personality trade-off by plotting PAS against DP for both models and datasets. Each arrow represents the shift from neutral to sensitive prompts. Arrows pointing leftward and upward denote improvements in both fairness and personality alignment. However, we observe rightward trends for most transitions, reflecting worsened demographic fairness despite better alignment scores.

Figure 3 illustrates the *fairness-personality trade-off* in movie recommendations generated by ChatGPT and DeepSeek under two types of prompts: *neutral* and *personality-sensitive*. The *neutral prompt*, designed to appeal to a general audience, results in a broad selection of

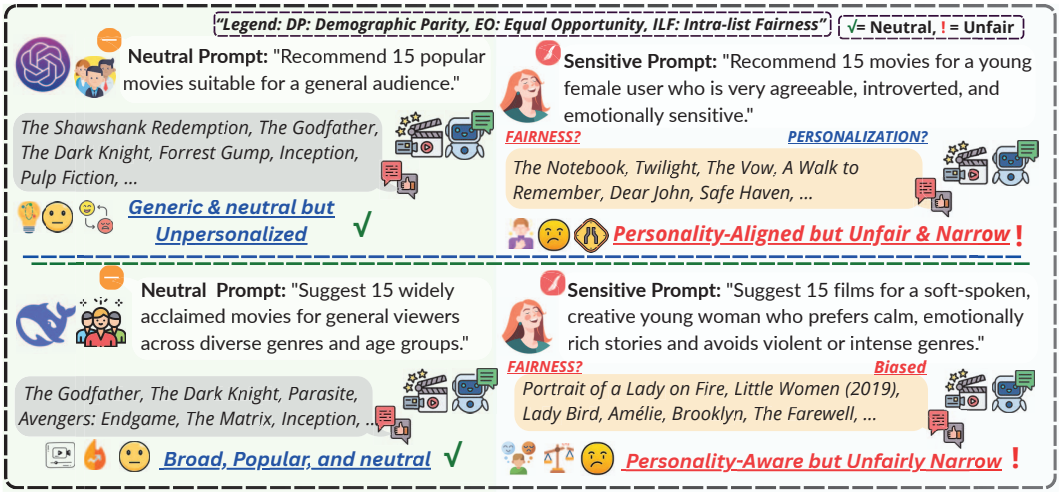

Figure 3. An illustration of the fairness-personality trade-off in LLM movie recommendations by ChatGPT and DeepSeek.

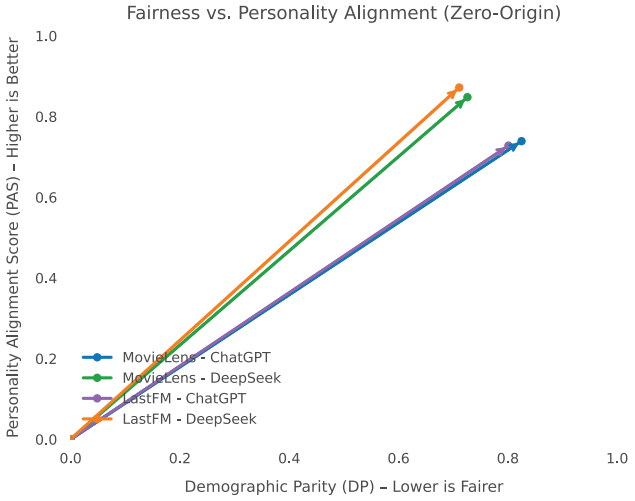

Figure 4. Trade-off between fairness (DP) and personality alignment (PAS). Arrows: rightward indicates increased disparity; upward indicates improved alignment.

movies, such as *The Shawshank Redemption*, *The Dark Knight*, and *Inception*. These recommendations are **Neutral** across demographic groups but *unpersonalized*, as they prioritize popularity over individual preferences. In contrast, the *personality-sensitive prompt* tailors recommendations based on specific user traits, such as *agreeableness*, *introversion*, and *emotional sensitivity*. This leads to a more *personalized* set of movies, including titles like *The Notebook*, *Lady Bird*, and *Little Women (2019)*, but these selections are **biased** toward particular genres, neglecting diversity and fairness. This trade-off between *fairness* and *personality alignment* highlights a crucial challenge in LLM-based recommender systems, where *personalization* often conflicts with *demographic fairness* and offering better alignment with user traits but often reflecting narrower genre preferences, reducing demographic neutrality. Personality-sensitive prompting enhances user alignment and intra-list diversity but introduces fairness risks. Designers must navigate this trade-off when deploying LLMs in recommendation pipelines.

4.3. RQ3: Model Comparison

To assess how different large language models (LLMs) balance personality alignment and fairness, we compare ChatGPT and DeepSeek across eight key metrics: Personality Alignment Score (PAS), Genre Preference Alignment (GPA), Demographic Parity (DP), Equal Opportunity (EO), Intra-List Fairness (ILF), Precision@15, Recall@15, and Jaccard@15. As shown in Table 4, DeepSeek achieves superior scores in PAS, ILF, Precision@15, and Recall@15 across both MovieLens and Last.fm datasets, indicating better alignment with user personality traits and stronger recommendation diversity and accuracy. Although ChatGPT performs marginally better on GPA and exhibits lower bias in some fairness metrics (e.g., DP and EO), DeepSeek's overall balance across all dimensions is stronger.

Figure 5 provides a visual summary of model performance under personality-sensitive prompts. DeepSeek consistently outperforms ChatGPT on six out of eight metrics, confirming its advantage in personalization-aware fairness. Additionally, We also illustrate this trade-off through qualitative prompt outputs in Figure 3. For example, personality-sensitive prompts for agreeable and emotionally sensitive users yield personalized yet demographically narrow recommendations such as *The Notebook* and *Lady Bird*. In contrast, neutral prompts return more diverse but generic titles like *Inception* and *The Dark Knight*, highlighting the personalization–fairness conflict.

Figure 5 graphically compares ChatGPT and DeepSeek

Table 4. Comparison of ChatGPT vs. DeepSeek on MovieLens and Last.fm datasets under personality-sensitive prompts. ↑: higher is better, ↓: lower is better.

Dataset	Metric	ChatGPT	DeepSeek
MovieLens	PAS ↑	0.739	0.848
	GPA ↑	0.407	0.336
	DP ↓	0.825	0.726
	EO ↓	0.952	0.901
	ILF ↑	0.310	0.968
	Prec@15 ↑	0.050	0.150
	Rec@15 ↑	0.015	0.040
	Jac@15 (N vs S) ↓	0.250	0.180
Last.fm	PAS ↑	0.728	0.872
	GPA ↑	0.421	0.352
	DP ↓	0.801	0.711
	EO ↓	0.935	0.884
	ILF ↑	0.289	0.932
	Prec@15 ↑	0.060	0.165
	Rec@15 ↑	0.020	0.045
	Jac@15 (N vs S) ↓	0.240	0.190

Note. PAS: Personality Alignment, GPA: Genre Preference Alignment, DP: Demographic Parity, EO: Equal Opportunity, ILF: Intra-list Fairness, Jac@15: Jaccard Similarity between Neutral and Sensitive prompts.

Table 5. PerFairX aggregated scores (FPx) for ChatGPT and DeepSeek across both datasets. Scores are computed using equal weights for all eight metrics: PAS, GPA, DP, EO, ILF, Precision@15, Recall@15, and Jaccard@15.

Dataset	Model	PerFairX Score
MovieLens	ChatGPT	1.994
MovieLens	DeepSeek	**2.895**
Last.fm	ChatGPT	2.022
Last.fm	DeepSeek	**2.961**

Note. The PerFairX score aggregates personality and fairness metrics into a single value. Higher scores reflect stronger overall alignment with fairness and personalization goals.

across eight core evaluation metrics under personality-sensitive prompting. Blue bars represent ChatGPT and green bars represent DeepSeek. As observed, DeepSeek significantly outperforms ChatGPT in most dimensions—particularly in personality alignment (PAS), intra-list diversity (ILF), and recommendation accuracy (Precision@15 and Recall@15). While ChatGPT maintains slightly lower disparity in demographic fairness metrics (DP and EO), its overall balance across the evaluation criteria is lower. To consolidate these multi-dimensional results, we apply our PerFairX aggregation formula to compute a unified scalar score for each model. As shown in Table 5, DeepSeek achieves higher PerFairX scores on both datasets—2.895 on MovieLens and 2.961 on Last.fm—surpassing ChatGPT's scores of 1.994 and

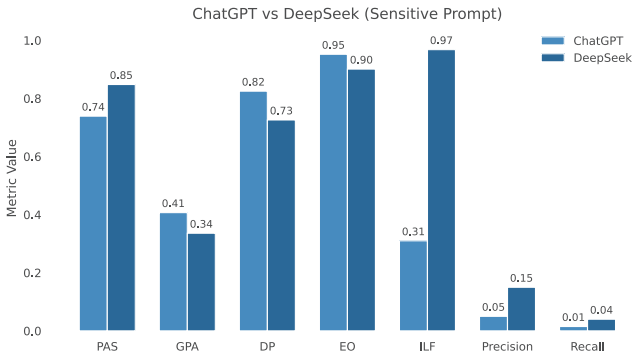

Figure 5. Comparison of ChatGPT vs. DeepSeek across seven metrics under personality-sensitive prompts. DeepSeek shows higher alignment (PAS), diversity (ILF), and accuracy (Precision/Recall), while ChatGPT exhibits lower disparity on some fairness metrics (DP, EO).

2.022, respectively. These results emphasize DeepSeek's greater capacity to jointly optimize personalization and fairness objectives in LLM-based recommendation systems. DeepSeek delivers more personalized, diverse, and accurate recommendations, while ChatGPT is slightly fairer on group-level metrics. Overall, DeepSeek provides a better trade-off, as reflected by the PerFairX scores.

5. Conclusion

With the rapid advancement of LLMs, their role in powering next-generation recommender systems has gained increasing attention [2, 4, 24, 44, 45]. In this study, we introduced PerFairX, a novel framework to evaluate fairness and personality alignment in LLM-based rec. Analysis on two datasets with state-of-the-art LLMs shows that personality-sensitive prompting enhances user trait alignment but often increases demographic disparities, highlighting a key trade-off between personalization and fairness. These findings reinforce recent observations that personalization and equity are not inherently compatible in large-scale recommenders [33, 39, 43]. Notably, DeepSeek demonstrates superior alignment and diversity but greater prompt sensitivity, while ChatGPT offers more stable and perceived-as-fair outputs. PerFairX provides a principled benchmark to evaluate and compare such trade-offs, promoting transparent, human-centered system design [32]. Future work will extend this evaluation to models like Gemini, Claude, and LLaMA, and investigate fairness-aware prompt tuning, dynamic personality modeling, and debiasing strategies [28]. Ultimately, this research contributes to growing efforts to harmonize fairness and personalization in LLM-based recommender systems [3, 25, 38, 42], contributing to the broader discourse on ethical AI in multimodal and continual learning environments LLM Rec, Behavior Modeling, Counterfactual Fairness, User-Centric Design.

6. Acknowledgments

This work was supported by the the Software Engineering Institute of Beihang University and Chinese Government Scholarship. Heartfelt gratitude to Prof. Dr. Xiaoli Lian for invaluable support, supervision, guidance.

References

[1] Roza Abolghasemi, Paal Engelstad, Enrique Herrera-Viedma, and Anis Yazidi. A personality-aware group recommendation system based on pairwise preferences. *Information Sciences*, 595:1–17, 2022. 3

[2] Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. Information retrieval meets large language models: a strategic report from chinese ir community. *AI Open*, 4:80–90, 2023. 8

[3] Gabrielle Alves, Dietmar Jannach, Rodrigo Ferrari De Souza, and Marcelo Garcia Manzato. User perception of fairness-calibrated recommendations. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 78–88, 2024. 8

[4] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1007–1014, 2023. 8

[5] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2212–2220, 2019. 1, 3

[6] Robin Burke. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093*, 2017. 1

[7] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024. 2

[8] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447, 2024. 2, 4

[9] Yashar Deldjoo. Fairevalllm. a comprehensive framework for benchmarking fairness in large language model recommender systems. *arXiv e-prints*, pages arXiv–2405, 2024. 2

[10] Yashar Deldjoo and Tommaso Di Noia. Cfairllm: Consumer fairness evaluation in large-language model recommender system. *arXiv preprint arXiv:2403.05668*, 2024. 2

[11] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction*, 34(1): 59–108, 2024. 2

[12] Karlijn Dinnissen and Christine Bauer. Fairness in music recommender systems: A stakeholder-centered mini review. *Frontiers in big Data*, 5:913608, 2022.

[13] Andres Ferraro, Michael D Ekstrand, and Christine Bauer. It's not you, it's me: the impact of choice models and ranking strategies on gender imbalance in music recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 884–889, 2024. 2

[14] Bruce Ferwerda, Emily Yang, Markus Schedl, and Marko Tkalcic. Personality and taxonomy preferences, and the influence of category choice on the user experience for music streaming services. *Multimedia tools and applications*, 78: 20157–20190, 2019. 5

[15] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50 (3):1097–1179, 2024. 2

[16] Julian Gómez and Paola Ramírez. Intelligent news advertisement: A prompt learning approach to personalization using large language models. *Eastern European Journal for Multidisciplinary Research*, 3(2):325–337, 2024. 2

[17] David J Hughes, Ioannis K Kratsiotis, Karen Niven, and David Holman. Personality traits and emotion regulation: A targeted review and recommendations. *Emotion*, 20(1): 63, 2020. 2

[18] Meng Jiang, Keqin Bao, Jizhi Zhang, Wenjie Wang, Zhengyi Yang, Fuli Feng, and Xiangnan He. Item-side fairness of large language model-based recommendation system. In *Proceedings of the ACM Web Conference 2024*, pages 4717–4726, 2024. 2, 4

[19] Elisabeth Lex and Markus Schedl. Psychology-informed recommender systems tutorial. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 714–717, 2022. 2

[20] Elisabeth Lex, Dominik Kowald, Paul Seitlinger, Thi Ngoc Trang Tran, Alexander Felfernig, Markus Schedl, et al. Psychology-informed recommender systems. *Foundations and trends® in information retrieval*, 15(2):134–242, 2021. 3

[21] Yijiang Li and Jun Wu. Improving prompt-based news recommendation with individual template and customized answer. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, page 3887–3891, New York, NY, USA, 2024. Association for Computing Machinery. 2

[22] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. Towards personalized fairness based on causal notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1054–1063, 2021. 2

[23] Wei Liu, Baisong Liu, Jiangcheng Qin, Xueyuan Zhang, Weiming Huang, and Yangyang Wang. Fairness identification of large language models in recommendation. *Scientific Reports*, 15(1):5516, 2025. 2

[24] Zheng Liu, Yujia Zhou, Yutao Zhu, Jianxun Lian, Chaozhuo Li, Zhicheng Dou, Defu Lian, and Jian-Yun Nie. Information retrieval meets large language models. In *Companion*

Proceedings of the ACM Web Conference 2024, pages 1586–1589, 2024. 8

[25] Sebastian Lubos, Thi Ngoc Trang Tran, Alexander Felfernig, Seda Polat Erdeniz, and Viet-Man Le. Llm-generated explanations for recommender systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 276–285, 2024. 8

[26] Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. Llm-rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780*, 2023. 1, 4

[27] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992. 5

[28] Helena Mihaljević, Ivana Müller, Katja Dill, Aysel Yollu-Tok, and Maximilian von Grafenstein. More or less discrimination? practical feasibility of fairness auditing of technologies for personnel selection. *AI & SOCIETY*, 39(5):2507–2523, 2024. 8

[29] What Is Data Mining. Data mining: Concepts and techniques. *Morgan Kaufinann*, 10(559-569):4, 2006. 4

[30] Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. Fairlens: Auditing black-box clinical decision support systems. *Information Processing & Management*, 58(5):102657, 2021. 2

[31] Chandan Kumar Sah, Lian Xiaoli, and Muhammad Mirajul Islam. Unveiling bias in fairness evaluations of large language models: A critical literature review of music and movie recommendation systems. *arXiv preprint arXiv:2401.04057*, 2024. 2

[32] Chandan Kumar Sah, Lian Xiaoli, Muhammad Mirajul Islam, and Md Kamrul Islam. Navigating the ai frontier: a critical literature review on integrating artificial intelligence into software engineering education. In *2024 36th International Conference on Software Engineering Education and Training (CSEE&T)*, pages 1–5. IEEE, 2024. 8

[33] Chandan Kumar Sah, Xiaoli Lian, Tony Xu, and Li Zhang. Faireval: Evaluating fairness in llm-based recommendations with personality awareness. *arXiv preprint arXiv:2504.07801*, 2025. 2, 8

[34] Christopher J Soto and Oliver P John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117, 2017. 2

[35] Zhaoxuan Tan and Meng Jiang. User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518*, 2023. 2

[36] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM conference on recommender systems*, pages 269–278, 2020. 2

[37] Antonela Tommasel. Fairness matters: A look at llm-generated group recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 993–998, 2024. 2

[38] Xiangmeng Wang, Qian Li, Dianer Yu, Qing Li, and Guandong Xu. Counterfactual explanation for fairness in recommendation. *ACM Transactions on Information Systems*, 42 (4):1–30, 2024. 8

[39] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023. 8

[40] Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. Joint multisided exposure fairness for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on research and development in information retrieval*, pages 703–714, 2022. 1, 2, 3

[41] Emre Yalcin and Alper Bilge. Popularity bias in personality perspective: An analysis of how personality traits expose individuals to the unfair recommendation. *Concurrency and Computation: Practice and Experience*, 35(9):e7647, 2023. 2

[42] Dayu Yang, Fumian Chen, and Hui Fang. Behavior alignment: a new perspective of evaluating llm-based conversational recommendation systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2286–2290, 2024. 8

[43] Hyunsik Yoo, Zhichen Zeng, Jian Kang, Ruizhong Qiu, David Zhou, Zhining Liu, Fei Wang, Charlie Xu, Eunice Chan, and Hanghang Tong. Ensuring user-side fairness in dynamic recommender systems. In *Proceedings of the ACM Web Conference 2024*, pages 3667–3678, 2024. 8

[44] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. Where to go next for recommender systems? id-vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2639–2649, 2023. 8

[45] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 993–999, 2023. 1, 2, 4, 8

[46] Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. Towards empathetic conversational recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 84–93, 2024. 2

[47] Yuying Zhao, Yu Wang, Yi Zhang, Pamela Wisniewski, Charu Aggarwal, and Tyler Derr. Leveraging opposite gender interaction ratio as a path towards fairness in online dating recommendations based on user sexual orientation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22547–22555, 2024. 2

[48] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021. 2